# EasySVM: A Visual Analysis Approach for Open-Box Support Vector Machines

Yuxin Ma, Wei Chen, Xiaohong Ma, Jiayi Xu, Xinxin Huang, Ross Maciejewski, and Anthony K.H. Tung

**Abstract**—

**Index Terms**—Support vector machines, rule extraction, visual classification, high-dimensional visualization, visual analysis.

## 1 Introduction

A support vector machine (SVM) [?] is a supervised learning method that is widely used in a variety of application areas, such as text analysis [?], computer vision [?] and bioinformatics [?], [?]. The SVM model is a discriminative model which tries to split the training data into the two classes by creating a *separating hyper-plane* at the place where the two classes are furthest apart. The class of a new data point is predicted by determining which side of the hyper-plane it laid on.

The main contributions include:

- An interactive visualization method for exploring data instances, SVM models and their relationships;
- A visual analysis approach for model comparison and selection, and;
- A visual rule extraction method that allows the user to extract rules that best interpret the models.

## 2 Related Work

The work presented in this paper is related to three broad topics: 1) support vector machines; 2) visual exploration of high-dimensional data, and; 3) visual classification.

### 2.1 Support Vector Machines

The SVM is currently regarded as state-of-the art in classification techniques [?], and studies have revealed that SVMs perform well when compared to other classification techniques [?]. This performance can be partly attributed to the use of non-linear kernels which unfortunately make it difficult to interpret the models. In addition, the production of the boundary function is often quite difficult. Work by Wahba et al. [?] explored the use of SVM functionals to produce classification boundaries, exploring tradeoffs between the size of the SVM functional and the smoothing parameters.

Rule extraction is an important component for the interpretation of SVMs or other classification techniques [?], [?]. Martens et al. [?] provided a comprehensive study on rule extraction of SVMs. These methods commonly employ an automatic optimization process and result in an axis-parallel representation. However, the targets and interests may vary according to the user and analysis tasks. It is likely that a visual analysis process can enable the user to explore both the input parameter space and the classification boundaries.

Unfortunately, there is little work that is dedicated to visualizing SVMs. Caragea et al. [?] applied a projection method to transform data instances onto 2-D space. The separating hyper-plane is sampled in the data space, and projected on to the 2-D plane. Hamel et al. [?] proposed to visualize data instances and SVM models with self-organized map (SOM). The work by Aragon et al. [?] utilized SVMs as part of a visual analysis system for astrophysics, but provided no general support for SVM exploration. As such, past visualization work has focused on generating static visualizations, and does not support interactive exploration.

### 2.2 Visual Exploration of High-Dimensional Data

One key challenge in opening up SVMs is the need for high-dimensional data exploration methods. Recent work in this area has utilized multi-dimensional projections to map data instances in high-dimensional data space to the low-dimensional (2-D) space. The key issue is how to explore

- *Yuxin Ma, Wei Chen, Xiaohong Ma, Jiayi Xu and Xinxin Huang are with the State Key Lab of CAD&CG, Zhejiang University.*
- *Ross Maciejewski is with Arizona State University.*
- *Anthony K.H. Tung is with National University of Singapore.*

*Manuscript received January 31, 2015; revised January 31, 2015.*

the underlying dataset with informative projections. Previous works, such as grand tour [?] and projection pursuit [?], generate a series of projections that allow the user to dynamically explore various lower dimensional projections of the data in a systematic way in order to find a preferred projection. Other exploration techniques facilitate the user with controls of the projection matrix (e.g., [?], [?]). Nam et al. [?] proposed a projection navigation interface, where the exploration and navigation activities are decomposed into five major tasks: sight identification, tour path planning, touring, looking around & zoom into detail, and orientation & localization. Additionally, an N-D touchpad polygon is provided to navigate in high-dimensional space by adjusting the combination of projection weights on each dimension.

Alternatively, high-dimensional data can be visualized with scatterplot matrix [?], parallel coordinates plot (PCP) [?], [?] and radar charts [?]. Previous work has also employed interactive exploration and navigation among scatterplots to fill the gap between projections and axes-based visualization techniques. Elmqvist et al. [?] presented an interactive method to support visualization and exploration of relations among different 2-D scatterplots in high-dimensional space. Similarly, 3D navigation [?] on the basis of rigid body rotation can be employed for viewing 3D scatterplot matrices. The 3-D rotation interaction improves the user's ability to perceive corresponding points in different scatterplots for comparisons.

### 2.3 Visual Classification

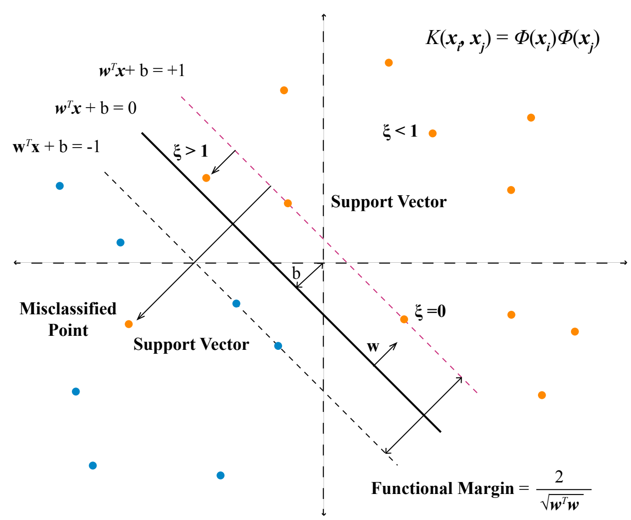
Some visual classification approaches, such as decision and rule-based classifiers [?], [?], employ so-called "white-box models", with which the detailed process is easy to understand. Teoh et al. [?] considered the process of building decision trees as a knowledge discovery method, and argued that visualization of the decision tree model can reveal valuable information in the data. Elzen et al. [?] presented a system for interactive construction and analysis of desicion trees with operations including growing, pruning, optimization and analysis.

Another category of works focuses on designing model-transparent frameworks where the user is allowed to provide training dataset and view the results. Thus, the low-level classification techniques can be directly employed without modification of the analytical process. Heimerl et al. [?] proposed to make a tight integration of the user into the labelling process and suggested an interactive binary classifier training approach for text analysis. Höferlin et al. [?] presented a system to build cascade of linear classifiers for image classification.

For open-box visual analysis approaches, one of the most similar works to our approach is from Tzeng et al. [?]. It combines several visualization designs for artificial neural networks to open the black box of underlying dependencies between the input and output data. Unlike our interactive visual analysis approach, their open-box scheme is limited to present a static visualization of the model structure and does not provide means of data exploration and interpretation of classification process.

## 3 An Introduction to SVM Classification

Given a set of training data points each with $m$ attributes and an associated class labels, the SVM attempts to separate these points by a $(m-1)$-dimensional hyperplane. In this section, we will briefly describe this process with the help of Figure 1.




Suppose that $\boldsymbol{x}_i \in \mathbb{R}^m, i = 1, 2, \ldots, n$ are $n$ training data instances in two different classes, and $y_i \in \{-1, +1\}, i = 1, 2, \ldots, n$ are their corresponding class labels. A linear support vector machine aims to construct a hyper-plane

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0 \tag{1}$$

in the $m$-dimensional data space $\mathbb{R}^m$ that has the largest distance to the nearest training data instances of each class (which is called "a functional margin").

Equation 1 can be solved by computing the following optimization problem:

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i \tag{2}$$
$$\text{subject to} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n.$$

where $C$ is a user-adjustable parameter to control the relative importance of maximizing the margin or satisfying the constraint of partitioning each training data instances into the correct half-space. The dual problem of Equation 2 is derived by introducing Lagrange multipliers $\alpha_i$:

$$\min \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i \tag{3}$$
$$\text{subject to} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, n.$$

Here, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) = \langle\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)\rangle$ is called the *kernel function*. For a linear SVM, its kernel is the dot product of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, i.e., $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$. Support vectors are those training data instances $\boldsymbol{x}_s$ whose corresponding Lagrange multipliers is above zero. Finally, the decision function for classifying a new data instance $\hat{\boldsymbol{x}}$ is:

$$\hat{y} = \text{sgn}(\sum_{i=1}^{n} y_i\alpha_i\hat{x}_i + b). \tag{4}$$

## 4 Overview

## 5 Open-box Visual Modeling for SVMs

### 5.1 Visualization of Data Instances and the SVM Model

(这部分与 workshop 版本基本一致，主要阐述 scatterplot view 使用的正交投影方法，以及视觉编码。需要重点强调的是我们选择正交投影方法的原因：if we can see separation between points within the scatterplot, then we can infer this separation exists in the high-dimensional space.)

### 5.2 Visual Exploration on Projected Scatterplot

(这里需要新增的是新的 manual control 设计。使用一个圆形控制盘来调整用户在高维空间中进行 trackball control 时每个维度上的权重。)

Exploration actions performed to extract knowledges from the dataset and SVM model:

- Training Data distributions, clusters or outliers（和 workshop 版本一样，新增一点说明是：用户可以根据 domain knowledge 等对训练数据集进行改动)
- Relations between model and data instances
- Optimization iterations analysis

## 6 Visual Model Comparison and Selection

### 6.1 Visualization and Visual Comparison of Multiple Models

在这里，我们将每个模型看做一个拥有多个属性（例如输入参数、训练数据集）的数据对象。我们主要分析这些属性组成的 property space 中的特征（比较类似 parameter space analysis）。

主要包括以下视图：

- View1(global): Revolution map：使用节点连接图方式表示已有模型的生成顺序和分支
- View2(global): Scatterplot(PCP) of properties：使用散点图（平行坐标）布局，横纵轴分别表示模型的某种属性，点（线）代表模型。可用于分析属性之间的关系。
- View3(local): Comparison view of specific models：将几个模型的 projected scatterplot 并排显示做对比。用于分析输入参数或训练数据具体如何影响模型训练结果之间的差异。

如何根据视图定位参数/训练数据集对模型的影响，即找到 Point of interest：

- Top-down startegy：从 View1 和 View2 开始，找到一些特征后选择几个模型，进入 view3 分析具体原因
- Bottom-up strategy：从 View3 出发，比较模型具体差异后将其放在 View2 和 View1 中，查看其在全局所有模型下所处的位置。

分析任务：

- Optimization：找到训练数据质量优、拟合结果较好地模型
- Partitioning(clustering of models)：从全局角度探索模型拟合结果主要有哪几种大的分类
- Outliers：哪些训练数据和输入参数的组合会产生特殊的模型
- Sensitivity analysis：参数/训练数据对模型拟合结果变化的敏感性

列出以上分析任务的原因：除了最根本的 Optimization 以外，visual analytics 的最大作用就是让用户获得更多的 insight，而不是简单地得到最优结果。

经过比较后，训练数据质量优、拟合结果较好的模型可被选择出来。

## 7 Visual Rule Extraction

## 8 Case Study

## 9 Discussion

## 10 Conclusion

## References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

PLACE
PHOTO
HERE

**Michael Shell** Biography text here.

**John Doe** Biography text here.

**Jane Doe** Biography text here.